

# Nonparametric Model Checking and Variable Selection

ADRIANO ZAMBOM AND MICHAEL AKRITAS

The Pennsylvania State University

December 21, 2011

## Abstract

Let  $\mathbf{X}$  be a  $d$  dimensional vector of covariates and  $Y$  be the response variable. Under the nonparametric model  $Y = m(\mathbf{X}) + \sigma(\mathbf{X})\epsilon$  we develop an ANOVA-type test for the null hypothesis that a particular coordinate of  $\mathbf{X}$  has no influence on the regression function. The asymptotic distribution of the test statistic, using residuals based on Nadaraya-Watson type kernel estimator and  $d \leq 4$ , is established under the null hypothesis and local alternatives. Simulations suggest that under a sparse model, the applicability of the test extends to arbitrary  $d$  through sufficient dimension reduction. Using p-values from this test, a variable selection method based on multiple testing ideas is proposed. The proposed test outperforms existing procedures, while additional simulations reveal that the proposed variable selection method performs competitively against well established procedures. A real data set is analyzed.

**Keywords:** Nonparametric regression; kernel regression; Lack-of-fit tests; Dimension reduction; Backward elimination.

**Acknowledgments:** This research was partially supported by CAPES/Fulbright grant 15087657 and NSF grant DMS-0805598.

# 1 Introduction

For a response variable  $Y$  and a  $d$  dimensional vector of the available covariates  $\mathbf{X}$  set  $m(\mathbf{X}) = E(Y|\mathbf{X})$ . The dual problems of testing for the significance of a particular covariate, and identification of the set of relevant covariates are very common both in applied research and in methodological investigations. Due to readily available software, these tasks are often performed under the assumption of a linear model,  $m(\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$ . Model checking fits naturally in the methodological context of hypothesis testing, while variable selection is typically addressed through minimization of a constrained or penalized objective function, such as Tibshirani's (1996) LASSO, Fan and Li's (2001) SCAD, Efron, Hastie, Johnstone and Tibshirani's (2004) least angle regression, Zou's (2006) adaptive LASSO, and Candes and Tao's (2007) Dantzig selector.

At a conceptual level, however, the two problems are intimately connected: dropping variable  $j$  from the model is equivalent to not rejecting the null hypothesis  $H_0^j : \beta_j = 0$ . Abramovich, Benjamini, Donoho and Johnstone (2006) bridged the methodological divide by showing that application of the false discovery rate (FDR) controlling procedure of Benjamini and Hochberg (1995) on  $p$  values resulting from testing each  $H_0^j$  can be translated into minimizing a model selection criterion of the form

$$\sum_{i=1}^n \left( Y_i - \sum_{j \in S} \hat{\beta}_j^S x_{ij} \right)^2 + \sigma^2 |S| \lambda, \quad (1)$$

where  $S$  is a subset of  $\{1, 2, \dots, d\}$  specifying the model,  $\hat{\beta}_i^S$  denotes the least squares estimator from fitting model  $S$ ,  $|S|$  is the cardinality of the subset  $S$ , and the penalty parameter  $\lambda$  depends both on  $d$  and  $|S|$ . This is similar to penalty parameters used in Tibshirani and Knight (1999), Birge and Massart (2001) and Foster and Stine (2004), which also depend on both  $d$  and  $|S|$ , and more flexible than the proposal in Donoho and Johnstone (1994) which uses  $\lambda$  depending only on  $d$ , as well as AIC and Mallow's  $C_p$  which use constant  $\lambda$ .

Working with orthogonal designs, Abramovich et al. (2006) showed that the global minimum of the penalized least squares (1) with the FDR penalty parameter is asymptotically minimax for  $\ell^r$  loss,  $0 < r \leq 2$ , simultaneously throughout a range of sparsity classes, provided the level  $q$  for the FDR is set to  $q < 0.5$ . Generalizations of this methodology to non-orthogonal designs differ mainly in the generation of the  $p$  values for testing  $H_0^j : \beta_j = 0$ , and the FDR method employed. Bunea, Wegkamp and Auguste (2006) use  $p$  values generated from the standardized regression coefficients resulting from fitting the full model and employ Benjamini and Yekutieli's (2001) method for controlling FDR under dependency, while Benjamini and Gavrilov (2009) use  $p$  values from a forward selection procedure where the  $i$ th stage  $p$ -to-enter is the  $i$ th stage constant in the multiple-stage FDR procedure in Benjamini, Krieger and Yekutieli (2006).

Model checking and variable selection procedures based on the assumption that the regression function is linear may fail to discern the relevance of covariates whose effect on  $m(\mathbf{x})$  is nonlinear. See Tables 2 and 5. Because of this, procedures for both model checking and variable selection have been developed under more general/flexible models. See, for example Li and Liang (2008), Wang and Xia (2008), Huang, Horowitz and Wei (2010), Storlie, Bondell, Reich and Zhang (2011), and references therein. However, the methodological approaches for variable selection under these more flexible models have been distinct from those of model checking.

This paper aims at showing that a suitably flexible and powerful nonparametric model checking procedure can be used to construct a competitive nonparametric variable selection procedure by exploiting the aforementioned conceptual connection between model checking and variable selection. Thus, this paper has two objectives. The first is to develop a procedure for testing whether a particular covariate contributes to the regression function in the context of a heteroscedastic nonparametric regression model. The second objective is to propose a variable selection procedure based on backward elimination using the Benjamini and

Yekutieli (2001) method applied on the  $d$  p-values resulting from testing for the significance of each covariate.

In Section 2, we formally describe the model and introduce the hypothesis, the test statistic and its asymptotic distribution under the null hypothesis and local alternatives. Section 3 presents the results of a simulation study where the performance of the proposed test statistic is compared to those of existing tests. In Section 4 the proposed variable selection procedure is described and compared, in simulation studies and a real data set, to well established variable selection methods.

## 2 Nonparametric Model Checking

### 2.1 The Hypothesis and the Test Statistic

Let  $Y$  be the response variable and  $\mathbf{X} = (X_1, \dots, X_d)$  the vector of available covariates. Set  $m(\mathbf{X}) = E(Y|\mathbf{X})$  for the regression function and define

$$\zeta = Y - m(\mathbf{X}). \quad (2)$$

From its definition it follows that  $E(\zeta|\mathbf{X}) = E(\zeta) = E(\zeta|X_j) = 0$ , for all  $j = 1, \dots, d$ . Setting  $\sigma^2(\mathbf{X}) = \text{Var}(\zeta|\mathbf{X})$ , we have the model

$$Y = m(\mathbf{X}) + \sigma(\mathbf{X})\epsilon, \quad (3)$$

where  $\epsilon$  is the standardized error  $\zeta$ . Based on a sample  $(Y_i, \mathbf{X}_i), i = 1, \dots, n$ , of iid observations from model (3), we will consider testing the hypothesis that the regression function does not depend on the  $j$ th covariate. For simplicity in notation we set  $\mathbf{X} = (\mathbf{X}_1, X_2)$ , where  $\mathbf{X}_1$  is of dimension  $(d-1)$  and  $X_2$  is univariate. Setting  $E(Y|\mathbf{X}_1) = m_1(\mathbf{X}_1)$  the hypothesis we will consider can be written as

$$H_0 : m(\mathbf{x}_1, x_2) = m_1(\mathbf{x}_1). \quad (4)$$

To fully appreciate the nature of this hypothesis, let  $F_{\mathbf{X}_1}, F_{X_2}$  denote the marginal distribution functions of  $\mathbf{X}_1, X_2$ , respectively, and consider the ANOVA-type decomposition

$$m(\mathbf{X}_1, X_2) = \mu + \tilde{m}_1(\mathbf{X}_1) + \tilde{m}_2(X_2) + \tilde{m}_{12}(\mathbf{X}_1, X_2), \quad (5)$$

where  $\mu = \int \int m(\mathbf{x}_1, x_2) dF_{\mathbf{X}_1}(\mathbf{x}_1) dF_{X_2}(x_2)$ ,  $\tilde{m}_1(\mathbf{x}_1) = \int m(\mathbf{x}_1, x_2) dF_{X_2}(x_2) - \mu$ ,  $\tilde{m}_2(x_2) = \int m(\mathbf{x}_1, x_2) dF_{\mathbf{X}_1}(\mathbf{x}_1) - \mu$ ,  $\tilde{m}_{12}(\mathbf{x}_1, x_2) = m(\mathbf{x}_1, x_2) - \mu - \tilde{m}_1(\mathbf{x}_1) - \tilde{m}_2(x_2)$ . Note that their definition implies  $\int \tilde{m}_1(\mathbf{x}_1) dF_{\mathbf{X}_1}(\mathbf{x}_1) = \int \tilde{m}_2(x_2) dF_{X_2}(x_2) = \int \tilde{m}_{12}(\mathbf{x}_1, x_2) dF_{\mathbf{X}_1}(\mathbf{x}_1) = \int \tilde{m}_{12}(\mathbf{x}_1, x_2) dF_{X_2}(x_2) = 0$ .

Under the null hypothesis (4) it further follows that

$$m_1(\mathbf{x}_1) = \mu + \tilde{m}_1(\mathbf{x}_1), \quad \tilde{m}_2(X_2) = \tilde{m}_{12}(\mathbf{X}_1, X_2) = 0.$$

In the case that  $\mathbf{X}_1, X_2$  are independent, we also have  $E(Y|X_2) = \mu$  under the null.

Let now  $m_1(\mathbf{X}_{1i}) = E(Y|\mathbf{X}_{1i})$ , as before, and define the null hypothesis residuals as

$$\xi_i = Y_i - m_1(\mathbf{X}_{1i}). \quad (6)$$

Since under the null hypothesis (4)  $m_1(\mathbf{X}_{1i}) = m(\mathbf{X}_i)$ , it follows that the null hypothesis residuals in (6) equal the residuals defined in (2) and thus

$$E(\xi_i | X_{2i}) = 0. \quad (7)$$

The idea for constructing the test statistic is to think of the  $\xi_i$  as data from a high-dimensional one-way ANOVA design with levels  $x_{2i}, i = 1, \dots, n$ . Because of (7), it follows that under the null hypothesis (4) there are no factor effects, and we can use the high-dimensional one-way ANOVA statistic of Akritas and Papadatos (2004) for testing (4), after dealing with two important details. First,  $m_1$  is not known and needs to be estimated. Second, the statistic of Akritas and Papadatos (2004) requires two or more observations per cell, but in regression designs we typically have only one response per covariate value.

To deal with the unknown  $m_1$  will use the Nadaraya-Watson kernel estimator,

$$\hat{m}_1(\mathbf{X}_{1i}) = \sum_{j=1}^n \left( \frac{K_{H_n}(\mathbf{X}_{1i} - \mathbf{X}_{1j})}{\sum_{l=1}^n K_{H_n}(\mathbf{X}_{1i} - \mathbf{X}_{1l})} \right) Y_j, \quad i = 1, \dots, n, \quad (8)$$

with  $K_{H_n}(\mathbf{x}) = |H_n|^{-1} K(H_n^{-1}\mathbf{x})$ , where  $K(\cdot)$  is a bounded  $(d-1)$ -variate kernel function of bounded variation and with bounded support, and  $H_n$  is a symmetric positive definite  $(d-1) \times (d-1)$  matrix called the bandwidth matrix. Set

$$\hat{\xi}_i = Y_i - \hat{m}_1(\mathbf{X}_{1i})$$

for the estimated null hypothesis residuals.

To deal with the requirement of more than one observation per cell we make use of smoothness conditions and augment each cell by including additional  $p-1$   $\hat{\xi}_\ell$ 's which correspond to the  $(p-1)/2$   $X_{2\ell}$  values that are nearest to  $X_{2i}$  on either side. To be specific, we consider the  $(\hat{\xi}_i, X_{2i})$ ,  $i = 1, \dots, n$ , arranged so that  $X_{2i_1} < X_{2i_2}$  whenever  $i_1 < i_2$ , and for each  $X_{2i}$ ,  $(p-1)/2 < i \leq n - (p-1)/2$ , define the nearest neighbor window  $W_i$  as

$$W_i = \left\{ j : |\hat{F}_{X_2}(X_{2j}) - \hat{F}_{X_2}(X_{2i})| \leq \frac{p-1}{2n} \right\}, \quad (9)$$

where  $\hat{F}_{X_2}$  is the empirical distribution function of  $X_2$ .  $W_i$  defines the augmented cell corresponding to  $X_{2i}$ . Note that the augmented cells are defined as sets of indices rather than as sets of  $\hat{\xi}_i$  values. The vector of  $(n-p+1)p$  constructed "observations" in the augmented one-way ANOVA design is

$$\hat{\boldsymbol{\xi}}_V = (\hat{\xi}_j, j \in W_{(p-1)/2+1}, \dots, \hat{\xi}_j, j \in W_{n-(p-1)/2})'. \quad (10)$$

Let  $MST = MST(\hat{\boldsymbol{\xi}}_V)$ ,  $MSE = MSE(\hat{\boldsymbol{\xi}}_V)$  denote the balanced one-way ANOVA mean squares due to treatment and error, respectively, computed on the data  $\hat{\boldsymbol{\xi}}_V$ . The proposed test statistic is based on

$$MST - MSE. \quad (11)$$

## 2.2 Asymptotic results

### 2.2.1 Asymptotic null distribution

**Theorem 2.1.** *Assume that the marginal densities  $f_{\mathbf{X}_1}$ ,  $f_{X_2}$  of  $\mathbf{X}_1$ ,  $X_2$ , respectively, are bounded away from zero, the second derivatives of  $f_{\mathbf{X}_1}$  and  $m_1(\mathbf{x})$  are uniformly continuous and bounded, that  $\sigma^2(\cdot, x_2) := E(\xi^2 | X_2 = x_2)$  is Lipschitz continuous,  $\sup_{\mathbf{x}} \sigma^2(\mathbf{x}) < \infty$ , and  $E(\epsilon_i^4) < \infty$ . Assume that the eigenvalues,  $\lambda_i$ ,  $i = 1, \dots, d-1$ , of the bandwidth matrix  $H_n$  defined in (8), converge to zero at the same rate and satisfy*

$$n\lambda_i^8 \rightarrow 0 \quad \text{and} \quad \frac{n\lambda_i^{2(d-1)}}{(\log n)^2} \rightarrow \infty, \quad i = 1, \dots, d-1. \quad (12)$$

Then, under  $H_0$  in (4), the asymptotic distribution of the test statistic in (11) is given by

$$n^{1/2}(MST - MSE) \xrightarrow{d} N(0, \frac{2p(2p-1)}{3(p-1)}\tau^2),$$

where  $\tau = \int \left[ \int \sigma^2(\mathbf{x}_1, x_2) f_{\mathbf{X}_1 | X_2 = x_2}(\mathbf{x}_1) d\mathbf{x}_1 \right]^2 f_{X_2}(x_2) dx_2$ .

An estimate of  $\tau^2$  can be obtained by modifying Rice's (1984) estimator as follows

$$\hat{\tau}^2 = \frac{1}{4(n-3)} \sum_{j=2}^{n-2} (\hat{\xi}_j - \hat{\xi}_{j-1})^2 (\hat{\xi}_{j+2} - \hat{\xi}_{j+1})^2. \quad (13)$$

The next subsection gives the asymptotic theory under local additive and under general local alternatives. As these limiting results show, the asymptotic mean of the test statistic  $MST - MSE$  is positive under alternatives. Thus, the test procedure rejects the null hypothesis for "large" values of the test statistic.

### 2.2.2 Asymptotics under local alternatives

The local additive alternatives and the general local alternatives are of the form

$$H_1^A : m(\mathbf{x}_1, x_2) = m_1(\mathbf{x}_1) + \rho_n \tilde{m}_2(x_2), \quad (14)$$

$$H_1^G : m(\mathbf{x}_1, x_2) = m_1(\mathbf{x}_1) + \rho_{1n} \tilde{m}_2(x_2) + \rho_{2n} \tilde{m}_{12}(\mathbf{x}_1, x_2), \quad (15)$$

where the functions  $\tilde{m}_2$ ,  $\tilde{m}_{12}$  satisfy  $E(\tilde{m}_2(X_2)) = 0 = E(\tilde{m}_{12}(\mathbf{x}_1, X_2))$  and  $\rho_n = \rho_{1n} = an^{-1/4}$ ,  $\rho_{2n} = bn^{-1/4}$ , for constants  $a, b$ .

**Theorem 2.2.** *Consider the notation and assumptions of Theorem 2.1. Moreover, assume that  $\tilde{m}_2(x)$  is Lipschitz continuous.*

1. **(Local Additive Alternatives)** *Then, under  $H_1^A$  in (14), as  $n \rightarrow \infty$ ,*

$$n^{1/2}(MST - MSE) \xrightarrow{d} N\left(a^2 p \text{Var}(\tilde{m}_2(X_2)), \frac{2p(2p-1)}{3(p-1)}\tau^2\right).$$

2. **(Local General Alternatives)** *Assume further that  $\tilde{m}_{12}(\mathbf{x}_1, x_2)$  is Lipschitz continuous on  $x_2$  uniformly on  $\mathbf{x}_1$ . Then, under  $H_1^G$  in (15), as  $n \rightarrow \infty$ ,*

$$n^{1/2}(MST - MSE) \xrightarrow{d} N\left(\mu^G, \frac{2p(2p-1)}{3(p-1)}\tau^2\right), \text{ where}$$

$\mu^G = pa^2 \text{Var}(\tilde{m}_2(X_2)) + pb^2 \text{Var}(\tilde{m}_{12}(\mathbf{X}_1, X_2)) + 2pab \text{Cov}(\tilde{m}_2(X_2), \tilde{m}_{12}(\mathbf{X}_1, X_2))$ . If  $a = b$  the formula simplifies to  $\mu^G = pa^2 \text{Var}(\tilde{m}_2(X_2) + \tilde{m}_{12}(\mathbf{X}_1, X_2))$ .

## 2.3 Practical considerations

### 2.3.1 Using other estimators of $m(\mathbf{x}_1)$

We conjecture that the asymptotic theory of the test statistic remains the same for a wide class of other nonparametric estimators of  $m_1$ , such as local polynomial estimators or, under an additive model, the backfitting estimator. Moreover, if one is willing to assume additional smoothness conditions then, since use local polynomial estimators yields faster rates of convergence (Stone, 1982), we conjecture that the asymptotic theory of the test statistic based on such estimators can include covariate dimensionality greater than the present 4. A similar comment applies if one is willing to assume an additive model.

An alternative version of the present kernel estimator incorporated in our simulations is a version of the estimator proposed by Newey (1994) and Linton and Nielsen (1995), and

further studied in Mammen, Linton and Nielsen (1999) and Horowitz and Mammen (2004), computed as

$$\hat{\tilde{m}}_1(\mathbf{x}_1) = \frac{1}{n} \sum_{i=1}^n \hat{m}(\mathbf{x}_1, X_{2i}),$$

with  $\hat{m}(\mathbf{x}_1, X_{2i})$  a Nadaraya-Watson kernel estimator of  $m(\mathbf{x}_1, x_2)$ . Under the null hypothesis this also estimates  $E(Y|\mathbf{x}_1)$ , but under the alternative it estimates (see decomposition (5))

$$\tilde{m}_1(\mathbf{x}_1) = \mu + \tilde{m}_1(\mathbf{x}_1).$$

Table 1: Rejection rates with alternative fitting methods

| Method              | $\theta = 0$ |      | $\theta = 1$ |      |      |
|---------------------|--------------|------|--------------|------|------|
|                     | $\gamma$     |      | $\gamma$     |      |      |
|                     | 0            | 1    | 2            | 3    | 4    |
| ANOVA-type (p=11)   | .053         | .119 | .196         | .292 | .390 |
| ANOVA-type (p=9)    | .052         | .121 | .191         | .296 | .388 |
| ANOVA-type 2 (p=11) | .054         | .150 | .218         | .315 | .440 |
| ANOVA-type 2 (p=9)  | .054         | .122 | .201         | .320 | .422 |

In contrast, under the alternative,  $\hat{m}_1(\mathbf{x}_1)$  estimates

$$m_1(\mathbf{x}_1) = \mu + \tilde{m}_1(\mathbf{x}_1) + E(\tilde{m}_2(X_2) + \tilde{m}_{12}(\mathbf{x}_1, X_2)|\mathbf{X}_1 = \mathbf{x}_1).$$

Thus, forming the residuals by  $\hat{\xi} = Y - \hat{m}_1(\mathbf{x}_1)$  inadvertently removes some of the effect of  $X_2$ . The simulations reported in Table 1 suggest that the test statistic using the residuals  $\hat{\xi} = Y - \hat{m}_1(\mathbf{x}_1)$  (ANOVA-type2 in the table) can have improved power against non-additive alternatives. In this table, the data are generated according to the model  $Y = X_1 + \theta X_2 + \gamma X_1 X_2 + \epsilon$ , with  $X_1, X_2$  independent  $U(0, 1)$  random variables and  $\epsilon \sim N(0, 3^2)$ . The reported rejection rates are based on 2000 simulation runs with  $n = 100$ . We conjecture that a similar alternative to the local polynomial estimator of  $m_1$  will have improved power against non-additive alternatives.

### 2.3.2 Using dimension reducing techniques

The conditions of Theorem 2.1 restrict  $d$  to be less than or equal to 4. However, under the assumption of a sparse model, the effects of the curse of dimensionality can be moderated through the use of dimension reduction techniques. In all simulations reported in Section 4.2 as well as the data analysis results of Section 4.3 we used the classical sliced inverse regression (SIR) dimension reduction method of Li (1991). Moreover, we employed a variable screening method prior to applying SIR. The variable screening consists of performing the marginal test of Wang, Akritas and Van Keilegom (2008) for the significance of each variable, and keeping those variables for which the p-value is less than 0.5.

## 3 Simulations: Model Checking Procedures

### 3.1 Brief literature review

Let  $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$  be the vector of  $d$  available predictors, with  $\mathbf{X}_1$  being  $d_1$ -dimensional. The problem of assessing the usefulness of  $\mathbf{X}_2$ , i.e. testing  $H_0 : m(\mathbf{x}_1, \mathbf{x}_2) = m_1(\mathbf{x}_1)$ , has been approached from different angles by many authors. The literature is extensive, so only a brief summary of some of the proposed ideas and the resulting test procedures is given below. For additional references see Hart (1997) and Racine, Hart and Li (2006).

One class of procedures is based on the idea that the null hypothesis residuals,  $\xi = Y - m_1(\mathbf{X}_1)$ , satisfy  $E(\xi|\mathbf{X}) = 0$  under  $H_0$  and  $E(\xi|\mathbf{X}) = m(\mathbf{X}) - m_1(\mathbf{X}_1)$  under the alternative. Thus,  $E(\xi E(\xi|\mathbf{X})|\mathbf{X}) = (m(\mathbf{X}) - m_1(\mathbf{X}_1))^2$  under the alternative and zero under the null. Using this idea, Fan and Li (1996) propose a test statistic based on estimating  $E[\xi f_1(\mathbf{X}_1)E(\xi f_1(\mathbf{X}_1)|\mathbf{X})f(\mathbf{X})]$  which equals  $E[(m(\mathbf{X}) - m_1(\mathbf{X}_1))^2 f_1(\mathbf{X})^2 f(\mathbf{X})]$  under the al-

ternative and zero under the null. Their test statistic is

$$\frac{1}{n} \sum_i [\tilde{\xi}_i \tilde{f}_1(\mathbf{X}_{1i})] \left[ \frac{1}{(n-1)h_n^d} \sum_{j \neq i} [\tilde{\xi}_j \tilde{f}_1(\mathbf{X}_{1j})] K\left(\frac{\mathbf{X}_i - \mathbf{X}_j}{h_n}\right) \right]$$

where  $\tilde{f}_1$  is the estimated density of  $\mathbf{X}_1$ ,  $\tilde{\xi}_i$  is the estimated residuals under the null hypothesis, and  $K$  is a kernel function. Fan and Li (1996) show that their test statistic is asymptotically normal under  $H_0$ . Lavergne and Vounq (2000) propose a test statistic based on different estimator of the same quantity as Fan and Li (1996), which is

$$\frac{(n-4)!}{n!} \sum_a (Y_i - Y_k)(Y_j - Y_l) L_n\left(\frac{\mathbf{X}_{1i} - \mathbf{X}_{1k}}{g_n}\right) L_n\left(\frac{\mathbf{X}_{1j} - \mathbf{X}_{1l}}{g_n}\right) K_n\left(\frac{\mathbf{X}_i - \mathbf{X}_j}{h_n}\right),$$

where  $\sum_a$  is the sum over all permutations of 4 distinct elements chosen from  $n$ ,  $L_n = g_n^{-d_1} L$  for a kernel  $L$  on  $\mathbf{R}^{d_1}$  and  $K_n = h_n^{-d} K$  for a kernel  $K$  on  $\mathbf{R}^d$ . Lavergne and Vounq (2000) show that their test statistic is also asymptotically normal under  $H_0$ .

A related class of procedures is based on direct estimation of  $E[(m(\mathbf{X}) - m_1(\mathbf{X}_1))^2 W(\mathbf{X})]$ , for some weight function  $W$ . See, for example, Aït-Sahalia, Bickel, and Stoker (2001). The use of such test statistics is complicated by the need to correct for their bias. See also the bootstrap-based procedure of Delgado and Manteiga (2001). Because of the computer intensive nature of bootstrap-based procedures, these are not included in our comparisons.

An additional class of test procedures uses alternatives based on Stone's (1985) additive model. We will consider the procedure proposed by Fan and Jiang (2005). This is based on Fan, Zhang, and Zhang's (2001) Generalized Likelihood Ratio Test (GLR), using a local polynomial approximation and the backfitting algorithm for estimating the additive components.

## 3.2 Numerical comparison

In this section we compare the proposed ANOVA-type and ANOVA-type2 statistics described in Section 2.3.1 to the statistics proposed by Lavergne and Vuong (2000) (LV in the

tables), Fan and Li (1996) (FL in the tables), and Fan and Jiang (2005) (GLR in the tables).

The data is generated according to the models (also used in Lavergne and Vuong, 2000)

$$Y = -X_1 + X_1^3 + f_j(X_2) + \epsilon, \quad j = 0, 1, 2, 3, 4, 5, 6, \quad (16)$$

where  $X_1, X_2$  are iid  $N(0, 1)$  and  $\epsilon \sim N(0, 4)$ . Here,  $f_0(x) = 0$ , which corresponds to the null hypothesis  $H_0 : m(x_1, x_2) = m(x_1)$ ;  $f_1(X_2) = .5X_2$ ,  $f_2(X_2) = X_2$  and  $f_3(X_2) = 2X_2$  give three linear alternatives, and  $f_4(X_2) = \sin(2\pi X_2)$ ,  $f_5(X_2) = \sin(\pi X_2)$ , and  $f_6(X_2) = \sin(2/3\pi X_2)$  give three non-linear alternatives. The kernel for the Nadaraya-Watson estimation of  $m(X_1)$  is the uniform on  $(-0.5, 0.5)$  density, and the bandwidth is selected through leave-one-out cross validation. The rejection rates shown in Table 2 for LV, FL, and F tests are taken from the simulation results reported in the LV paper (based on 2000 runs). It is important to note that, in each simulation setting, the LV paper reports several rejection rates for the LV and FL tests, each corresponding to different values of smoothing parameters. Since the best performing constants are different for different simulation settings, the rejection rates reported in Table 2 represent a) the most accurate alpha level achieved over all constants, and b) the best power achieved overall constants for each alternative. For comparison purposes, the rejection rates for the ANOVA-type tests and the GLR test are also based on 2000 simulation runs.

As expected, the F test achieved the best results for the three linear alternatives and the worse results for the three non-linear alternatives. The GLR test has higher power than the ANOVA-type tests against linear alternatives (which is partly explained by the fact it is based on normal likelihood), but is much less powerful against the first of the non-linear alternatives. As the non-linearity decreases ( $f_5$  and  $f_6$ ) the power of the GLR test improves.

The GLR test is designed for additive models, which is exactly the simulation setting of Table 2. Under non-additive alternatives, however, it can perform poorly as indicated by the simulations reported in the first part of Table 3. These simulations use sample size  $n = 200$

Table 2: Rejection rates under  $H_0$ , linear and non-linear alternatives

| n   | test                | linear |       |       |       | sine  |       |       |
|-----|---------------------|--------|-------|-------|-------|-------|-------|-------|
|     |                     | $f_0$  | $f_1$ | $f_2$ | $f_3$ | $f_4$ | $f_5$ | $f_6$ |
| 100 | LV                  | .041   | .098  | .482  | .991  | .182  | .266  | .319  |
|     | FL                  | .021   | .051  | .271  | .970  | .126  | .168  | .187  |
|     | ANOVA-type (p = 9)  | .052   | .218  | .79   | .999  | .423  | .523  | .535  |
|     | ANOVA-type (p = 7)  | .056   | .244  | .780  | .999  | .432  | .527  | .551  |
|     | ANOVA-type2 (p = 9) | .065   | .275  | .831  | 1     | .453  | .598  | .600  |
|     | GLRT                | .044   | .365  | .951  | 1     | .123  | .497  | .645  |
|     | F-test              | .051   | .695  | .997  | 1     | .046  | .055  | .222  |
| 200 | LV                  | .054   | .208  | .875  | 1     | .386  | .540  | .678  |
|     | FL                  | .025   | .083  | .695  | 1     | .289  | .395  | .471  |
|     | ANOVA-type (p = 9)  | .055   | .374  | .95   | .999  | .73   | .778  | .788  |
|     | ANOVA-type (p = 7)  | .051   | .376  | .979  | 1     | .746  | .820  | .820  |
|     | ANOVA-type2 (p = 9) | .069   | .487  | .999  | 1     | .821  | .882  | .884  |
|     | GLRT                | .036   | .656  | 1     | 1     | .188  | .877  | .936  |
|     | F-test              | .052   | .931  | 1     | 1     | .051  | .053  | .340  |

with data generated from the model  $Y = X_1^{X_2}(1+\theta X_3) + \frac{X_2^{(1+\theta X_3)}}{X_2} + \epsilon$ , where  $\epsilon \sim N(0, 0.1)$ , and  $X_1, X_2, X_3$  are i.i.d.  $U(0.5, 2.5)$ . The hypothesis tested is that  $m(X_1, X_2, X_3) = m_1(X_1, X_2)$ . The residuals for the ANOVA-type test in the first part of Table 3 are based on a Nadaraya-Watson fit with kernel the uniform on  $(-0.5, 0.5) \times (-0.5, 0.5)$  density and the common bandwidth selected through leave-one-out cross validation.

Table 3: Rejection rates for non-additive and heteroscedastic models

| test                   | non-additive alternatives |      |      |      |      | heteroscedastic alternatives |       |      |      |      |
|------------------------|---------------------------|------|------|------|------|------------------------------|-------|------|------|------|
|                        | $\theta$                  |      |      |      |      | $\theta$                     |       |      |      |      |
|                        | 0                         | 0.02 | .04  | .06  | .08  | 0                            | 0.025 | .05  | .1   | .2   |
| ANOVA-type ( $p = 9$ ) | .052                      | .176 | .609 | .940 | .994 | .053                         | .067  | .124 | .485 | .998 |
| GLR                    | .048                      | .082 | .110 | .189 | .304 | .465                         | .511  | .624 | .908 | 1    |

Finally, it should be mentioned that the GLR test does not maintain its level under heteroscedasticity. In simulations, reported in the second part of Table 3, under the additive but heteroscedastic model  $Y = X_1^2 + \theta \cos(\pi X_2) + X_2 \epsilon$ ,  $X_1, X_2$  i.i.d.  $N(0, 1)$ ,  $\epsilon \sim N(0, 0.5)$ , using sample size  $n = 200$ , the GLR test is very liberal while the ANOVA-type test maintains an

accurate level.

## 4 From Model Checking to Variable Selection

### 4.1 The proposed procedure

In this section we will assume a sparse regression model in the sense that there exists a subset of indices  $I_0 = \{j_1, \dots, j_{d_0}\} \subset \{1, \dots, d\}$  such that only the covariates  $X_j$  with  $j \in I_0$  influence the regression function. Moreover, we will assume the dimension reduction model of Li (1991), i.e.  $m(\mathbf{x}) = g(\mathbf{B}\mathbf{x})$ , where  $\mathbf{B}$  is a  $K \times d$  matrix. In this context we will describe the following variable selection procedure using backward elimination based on the Benjamini and Yekutieli (2001) method for controlling the false discovery rate (FDR):

1. Apply the variable screening procedure described in Section 2.3.2. With a slight abuse of notation, the vector of the remaining covariates and its dimension will be denoted by  $\mathbf{x}$  and  $d$ .
2. Use SIR to obtain the estimator  $\hat{\mathbf{B}}$ .
3. Obtain p-values from testing each of the hypotheses:

$$H_0^j : m(\mathbf{x}) = m_1(\mathbf{x}_{(-j)}), \quad j = 1, \dots, d, \quad (17)$$

where  $\mathbf{x}_{(-j)} = (x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_d)$ :

- (a) Compute the test statistic (see Theorem 2.1)

$$z_j = \sqrt{n}(MST_j - MSE_j) / \sqrt{\frac{2p(2p-1)}{3(p-1)} \hat{\tau}_j^2}$$

using residuals formed by a kernel estimator on the variables  $\hat{\mathbf{B}}_{(-j)}\mathbf{x}_{(-j)}$ , where  $\hat{\mathbf{B}}_{(-j)}$  is the  $K \times (d-1)$  matrix obtained by omitting the  $j$ th column of  $\hat{\mathbf{B}}$ .

- (b) Compute the p-value for  $H_0^j$  as  $\pi_j = 1 - \Phi(z_j)$ .

4. Compute

$$k = \max \left\{ j : \pi_{(j)} \leq \frac{i}{d} \frac{\alpha}{\sum_{l=1}^d l^{-1}} \right\} \quad (18)$$

for a choice of level  $\alpha$ , where  $\pi_{(1)}, \dots, \pi_{(d)}$  are the ordered p-values. If  $k = d$  stop and retain all variables. If  $k < d$

- (a) update  $\mathbf{x}$  by eliminating the covariate corresponding to  $\pi_{(d)}$ ,
- (b) update  $\hat{\mathbf{B}}$  by eliminating the column corresponding to the deleted variable, and
- (c) proceed to the next step.

5. Repeat steps 3 and 3b, with the updated  $vx$  and  $\hat{\mathbf{B}}$ .

**Remarks.** 1) Another approach for constructing a variable selection procedure is to use a single application of the Benjamini and Yekutieli (2001) method for controlling the false discovery rate (FDR). This is similar to one of the two procedures proposed in Bunea et al. (2006). However, this did not perform well in simulations and is not recommended. A backward elimination approach was Li, Cook and Nachtsheim (2005), but they did not use multiple testing ideas.

2) Based on our simulation results, the variable screening part (Step 1) of the variable selection procedure does not improve the performance. However, it was included in the simulations as it reduces the computational time.

## 4.2 Simulations: Variable selection procedures

Because the ANOVA-type2 method (see Section 2.3.1) is computationally more intensive, we used only the proposed variable selection method using the ANOVA-type test described in Section 2.1, with ANOVA cell sizes of 5 (when  $n = 40$ ), 7, and 9 (when  $n = 110$ ). The parameter  $\alpha$  was set to 0.07 in Table 4 (so the FDR is controlled at level  $(25 - 5)0.07/25 = 0.056$ ), and  $\alpha = 0.06$  in Table 5 (so the FDR is controlled at levels 0.052 and 0.045). These procedures are compared with LASSO, SCAD, adaptive LASSO, the FDR-based variable

selection method proposed by Bunea, Wegkamp and Auguste (2005) (BWA in the tables), and a version of the BWA procedure which uses backward elimination (BWA+BE in the tables). The comparison criterion is the mean number of correctly and incorrectly excluded variables. All comparisons are based on 2000 simulated data sets.

For LASSO we found that the R code in <http://cran.r-project.org/web/packages/glmnet/index.html>, with the `lambda.lse` option for selecting  $\lambda$ , gave the best results; for adaptive LASSO we used the R code from <http://www4.stat.ncsu.edu/~boos/var.select/lasso.adaptive.html>; for SCAD we used the function `scadglm` of the package SIS in R.

In Table 4, data sets of size  $n = 110$  were generated from the linear model  $Y = \beta^T \mathbf{X} + \epsilon$ , where  $\epsilon \sim N(0, 3^2)$ , the dimension of  $\mathbf{X}$  is  $d = 25$ , and

$$\beta^T = (3, 1.5, 0, 0, 2, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0).$$

The covariates are generated from a multivariate normal distribution with marginal means zero and covariances as shown in the table. It is seen that the proposed nonparametric variable selection procedures correctly exclude, on average, about 19.5 out of the 20 non-significant predictors. This is about as good as the procedures designed for linear models. The proposed procedures incorrectly exclude, on average, about 0.5 of the 5 significant predictors, which is more than the other procedures (with the exception of BWA).

| Table 4: Comparisons using a linear model: $d = 25$ , $n = 110$ |              |           |                          |           |
|---|--------------|-----------|--------------------------|-----------|
| test  | $\Sigma = I$ |           | $\Sigma = (0.5^{ i-j })$ |           |
|   | correct      | incorrect | correct                  | incorrect |
| SCAD  | 19.48        | .026      | 19.37                    | .023      |
| LASSO   | 18.29        | .005      | 18.28                    | .004      |
| Adaptive LASSO  | 19.28        | .005      | 19.26                    | .025      |
| BWA   | 19.99        | 1.02      | 19.97                    | 1.41      |
| BWA+BE  | 19.55        | .001      | 19.49                    | .041      |
| ANOVA-type(p=7)   | 19.46        | .63       | 19.30                    | .44       |
| ANOVA-type(p=9)   | 19.52        | .65       | 19.40                    | .36       |

In Table 5, data sets of size  $n = 40$  were generated from the models  $Y = g_\ell(\mathbf{X}) + \epsilon$ ,

$\ell = 1, 2$ , where  $\epsilon \sim N(0, 0.3^2)$ , the dimension of  $\mathbf{X}$  is  $d = 8$ , and

$$g_1(\mathbf{x}) = \sin(\pi x_1), \quad g_2(\mathbf{x}) = \sin(3/4\pi x_1) - 3\Phi(-|x_5|^3).$$

The covariates are generated as normal with marginal means zero and covariance matrix

Table 5: Comparisons using nonlinear models:  $d = 8$ ,  $n = 40$

| test            | $g_1$   |           | $g_2$   |           |
|-----------------|---------|-----------|---------|-----------|
|                 | correct | incorrect | correct | incorrect |
| SCAD            | 6.74    | .96       | 5.71    | 1.79      |
| LASSO           | 6.59    | .92       | 5.72    | 1.80      |
| Adaptive LASSO  | 6.65    | .95       | 5.62    | 1.73      |
| BWA             | 6.99    | 1         | 5.99    | 1.99      |
| BWA+BE          | 6.65    | .94       | 5.70    | 1.75      |
| ANOVA-type(p=7) | 6.21    | 0.001     | 5.75    | .11       |
| ANOVA-type(p=5) | 6.39    | 0.001     | 5.71    | .08       |

$\Sigma = (0.5^{|i-j|})$ . It is seen that the linear model based procedures fail to select the significant predictor(s) almost always. On the other hand, the proposed procedures always select the one relevant predictor under model  $g_1$ , and exclude incorrectly, on average, about 0.08 out of the two important predictors under model  $g_2$ .

### 4.3 Real Data Example: Body Fat Dataset

The Body Fat data is supplied by Dr. A. Garth Fisher for non-commercial purposes, and it can be found at "<http://lib.stat.cmu.edu/datasets/bodyfat>". The data set contains measurements of percent body fat (using Siri's (1956) method), Age (years), Weight (lbs), Height (inches), circumferences of Neck (cm), Chest (cm), Abdomen (cm), Hip (cm), Thigh (cm), Knee (cm), Ankle (cm), Biceps (cm), Forearm (cm) and Wrist (cm), from 252 men. The response variable is the percentage of body fat.

We compare the results of SCAD, LASSO, Adaptive LASSO and BWA with backward elimination to the ANOVA-type procedure with variable screening and SIR, as described in Section 4.1. Table 4.3 shows the results for LASSO, SCAD, Adaptive LASSO and BWA. It

Table 6: Results for LASSO, Adaptive LASSO, SCAD, BWA

| Predictor | LASSO   | Adpt. LASSO | SCAD    | BWA     |
|-----------|---------|-------------|---------|---------|
| Age       | .06499  | 0           | .001061 | 0       |
| Weight    | 0       | -.09511     | -.11688 | -.1356  |
| Height    | -.1591  | 0           | -.05818 | 0       |
| Neck      | -.2579  | 0           | 0       | 0       |
| Chest     | 0       | 0           | 0       | 0       |
| Abdomen   | .7079   | .9113       | .9052   | .9958   |
| Hip       | 0       | 0           | 0       | 0       |
| Thigh     | 0       | 0           | 0       | 0       |
| Knee      | 0       | 0           | 0       | 0       |
| Ankle     | 0       | 0           | 0       | 0       |
| Biceps    | 0       | 0           | 0       | 0       |
| Forearm   | .21756  | 0           | 0       | .4729   |
| Wrist     | -1.5353 | -.9871      | 0       | -1.5056 |

is seen that Weight and Abdomen are selected by all except LASSO. This can be explained by the fact that LASSO does not perform well in the presence of highly correlated variables, which is the case with this data set. The Adaptive LASSO and BWA give almost the same results but differ considerably from those of SCAD.

For the ANOVA-type method we used SIR with the number of slices ranging from 2 to 100. Abdomen, Weight, Biceps and Knee were selected with 99, 87, 88 and 23 of the 99 different numbers of slices, respectively. All other variables were selected less than 15 times. On the basis of these results we recommend a model based on Abdomen, Weight and Biceps.

As an explanation of the fact that Biceps was not selected by any of the other methods, we investigated possible violations of the modeling assumptions on which they are based. Marginal plots of the response versus each of the important variables reveal both heteroscedasticity and nonlinearity. Moreover, the 99 applications of SIR yielded more than one linear combination (i.e.  $K > 1$ ) 50 times. To put this number into perspective, we generated a single set of responses, using the same covariate values with coefficients those from Adaptive LASSO and normal errors using the residual variance. Application of SIR with the number of slices ranging from 2 to 100 on this data set yielded  $K = 1$  92 out of the

99 times. This casts serious doubts on the validity of the assumption of a linear model.

# Appendix

## A Auxiliary Results

**Lemma A.1.** *Let  $X_1, \dots, X_n$  be iid $[F]$ , and let  $\hat{F}_n(x)$  be the corresponding empirical distribution function. Then, for any constant  $c$ ,*

$$\sup_{x_i, x_j} \left\{ |F(x_i) - F(x_j)| I \left[ |\hat{F}_n(x_i) - \hat{F}_n(x_j)| \leq \frac{c}{n} \right] \right\} = O_p \left( \frac{1}{\sqrt{n}} \right).$$

*Proof.* By the Dvoretzky, Kiefer and Wolfowitz (1956) theorem, we have that  $\forall \epsilon \geq 0$ ,

$$P \left( \sup_x |\hat{F}_n(x) - F(x)| \geq \epsilon \right) \leq C e^{-2n\epsilon^2}.$$

Therefore,  $|\hat{F}_n(x) - F(x)| = O_p \left( \frac{1}{\sqrt{n}} \right)$  uniformly on  $x$ . Hence, writing

$$|F(x_i) - F(x_j)| = |F(x_i) - \hat{F}_n(x_i) + \hat{F}_n(x_i) - F(x_j) + \hat{F}_n(x_j) - \hat{F}_n(x_j)|,$$

it follows that  $\sup_{x_i, x_j} \left\{ |F(x_i) - F(x_j)| I \left[ |\hat{F}_n(x_i) - \hat{F}_n(x_j)| \leq c/n \right] \right\}$  is less than or equal to

$$\begin{aligned} & \sup_{x_i, x_j} \left\{ |F(x_i) - \hat{F}_n(x_i)| + |\hat{F}_n(x_j) - F(x_j)| \right\} \\ & + \sup_{x_i, x_j} \left\{ |\hat{F}_n(x_i) - \hat{F}_n(x_j)| \right\} I \left[ |\hat{F}_n(x_i) - \hat{F}_n(x_j)| \leq \frac{c}{n} \right] \\ & = O_p \left( \frac{1}{\sqrt{n}} \right) + O_p \left( \frac{1}{\sqrt{n}} \right) + O_p \left( \frac{1}{n} \right). \end{aligned}$$

This completes the proof of the lemma. □

**Lemma A.2.** *With  $W_i$  be defined in (9), and any Lipschitz continuous function  $g(x)$ ,*

$$\frac{1}{p} \sum_{j=1}^n g(x_{2j}) I(j \in W_i) - g(x_{2i}) = O_p \left( \frac{1}{\sqrt{n}} \right),$$

*uniformly in  $i = 1, \dots, n$ .*

*Proof.* First note that by the Lipschitz continuity and the Mean Value Theorem we have

$$|g(x_{2j}) - g(x_{2i})| \leq M|x_{2j} - x_{2i}| \leq M|F_{X_2}(x_{2j}) - F_{X_2}(x_{2i})|/f_{X_2}(\tilde{x}_{ij}),$$

for some constant  $M$ , where  $\tilde{x}_{ij}$  is between  $x_{2j}$  and  $x_{2i}$ . Thus,

$$\begin{aligned} \left| \frac{1}{p} \sum_{j=1}^n g(x_{2j}) I(j \in W_i) - g(x_{2i}) \right| &\leq \frac{1}{p} \sum_{j=1}^n |g(x_{2j}) - g(x_{2i})| I \left[ |\hat{F}_{X_2}(x_{2i}) - \hat{F}_{X_2}(x_{2j})| \leq \frac{p-1}{2n} \right] \\ &\leq \frac{M}{p} \sum_{j=1}^n \frac{|F_{X_2}(x_{2j}) - F_{X_2}(x_{2i})|}{f_{X_2}(\tilde{x}_{ij})} I \left[ |\hat{F}_{X_2}(x_{2i}) - \hat{F}_{X_2}(x_{2j})| \leq \frac{p-1}{2n} \right] = O_p \left( \frac{1}{\sqrt{n}} \right), \end{aligned}$$

where the last equality follows from Lemma A.1 and the assumption that  $f_{X_2}$  remains bounded away from zero.  $\square$

As in Wang, Akritas and Van Keilegom (2008), MST-MSE given in (11) can be written as a quadratic form  $\hat{\boldsymbol{\xi}}_V' A \hat{\boldsymbol{\xi}}_V$ , where

$$A = \frac{np-1}{n(n-1)p(p-1)} \oplus_{i=1}^n \mathbf{J}_p - \frac{1}{n(n-1)p} \mathbf{J}_{np} - \frac{1}{n(p-1)} \mathbf{I}_{np}, \quad (19)$$

where  $\mathbf{I}_d$  is a identity matrix of dimension  $d$ ,  $\mathbf{J}_d$  is a  $d \times d$  matrix of 1's and  $\oplus$  is the Kronecker sum or direct sum. Using arguments similar to those used in the proof of Lemma 3.1 in Wang, Akritas and Van Keilegom (2008), it can be shown that if  $\sigma^2(\cdot, x_2)$ , defined in Theorem 2.1, is Lipschitz continuous and  $E(\epsilon_i^4) < \infty$  then, under  $H_0$  and as  $n \rightarrow \infty$ ,

$$n^{1/2} [\boldsymbol{\xi}_V' A \boldsymbol{\xi}_V - \boldsymbol{\xi}_V' A_d \boldsymbol{\xi}_V] \xrightarrow{p} 0, \quad (20)$$

where  $A_d = \text{diag}\{B_1, \dots, B_n\}$ , with  $B_i = \frac{1}{n(p-1)} [\mathbf{J}_p - \mathbf{I}_p]$ .

**Lemma A.3.** *For a symmetric, positive definite bandwidth matrix  $H_n$ , define the norm  $\|H_n\|$  to be the maximum of its eigenvalues. Then we have*

$$\sum_{j_2=1}^n w(\mathbf{X}_{1i}, \mathbf{X}_{1j_2}) \|\mathbf{X}_{1j_2} - \mathbf{X}_{1i}\| = O(\|H_n^{1/2}\|).$$

*Proof.* Let  $b$  be such that  $K(\mathbf{x}) = K(\mathbf{x})I(\|\mathbf{x}\| \leq \sqrt{d-1}b)$ . Such a  $b$  exists by the assumption that the density  $K$  has bounded support. Thus,

$$\begin{aligned} & K(H_n^{-1/2}(\mathbf{X}_{1i} - \mathbf{X}_{1j})) \|\mathbf{X}_{1j} - \mathbf{X}_{1i}\| = K(H_n^{-1/2}(\mathbf{X}_{1i} - \mathbf{X}_{1j})) \|H_n^{1/2}H_n^{-1/2}(vX_{1i} - \mathbf{X}_{1j})\| \\ & \leq K(H_n^{-1/2}(\mathbf{X}_{1i} - \mathbf{X}_{1j})) \|H_n^{1/2}\| \|H_n^{-1/2}(vX_{1i} - \mathbf{X}_{1j})\| \\ & \leq K(H_n^{-1/2}(\mathbf{X}_{1i} - \mathbf{X}_{1j})) \|H_n^{1/2}\| \sqrt{d-1}b. \end{aligned}$$

The statement of the lemma follows from the above.  $\square$

## B Proofs of Theorems

***Proof of Theorem 2.1.*** Under  $H_0$  in (4) we write

$$\begin{aligned} \hat{\xi}_i &= Y_i - \hat{m}_1(\mathbf{X}_{1i}) + m_1(\mathbf{X}_{1i}) - m_1(\mathbf{X}_{1i}) = \xi_i - (\hat{m}_1(\mathbf{X}_{1i}) - m_1(\mathbf{X}_{1i})) \\ &= \xi_i - \Delta_{m_1}(\mathbf{X}_{1i}), \end{aligned}$$

where  $\Delta_{m_1}(\mathbf{X}_{1i})$  is defined implicitly in the above relation. Thus,  $\hat{\xi}_V$  of relation (10) is decomposed as  $\hat{\xi}_V = \xi_V - \Delta_{m_1V}$ , where  $\xi_V$  and  $\Delta_{m_1V}$  are defined as in (10) but using  $\xi_i$  and  $\Delta_{m_1}(\mathbf{X}_{1i})$ , respectively, instead of  $\hat{\xi}_i$ . Thus  $\sqrt{n}(\text{MST} - \text{MSE})$  can be written as

$$\sqrt{n}\hat{\xi}_V' A \hat{\xi}_V = \sqrt{n}\xi_V' A \xi_V - \sqrt{n}2\xi_V' A \Delta_{m_1V} + \sqrt{n}\Delta_{m_1V}' A \Delta_{m_1V}, \quad (21)$$

where the matrix  $A$  is defined in (19). The asymptotic normality of  $\sqrt{n}\xi_V' A \xi_V$  follows by arguments similar to those used in Theorem 3.2 of Wang, Akritas and VanKeilegom (2008). It remains to derive its asymptotic variance and to show that the other two terms in (21) converge to zero in probability. Using (20) it suffices to find the asymptotic variance of  $\sqrt{n}\xi_V' A_d \xi_V$ . Since  $E(\xi_V' A_d \xi_V) = 0$  its variance equals  $E[(\sqrt{n}\xi_V' A_d \xi_V)^2]$ . To find this we first evaluate its conditional expectation,  $E[(\sqrt{n}\xi_V' A_d \xi_V)^2 | \{X_{2j}\}_{j=1}^n]$ , given  $X_{21}, \dots, X_{2n}$ . Recalling the notation  $\sigma^2(\cdot, x_2) = E(\xi^2 | X_2 = x_2)$ , we have

$$\begin{aligned}
& \frac{1}{n(p-1)^2} \sum_{i_1, i_2}^n \sum_{j_1 \neq l_1}^n \sum_{j_2 \neq l_2}^n E(\xi_{j_1} \xi_{l_1} \xi_{j_2} \xi_{l_2} | \{X_{2j}\}_{j=1}^n) I(j_s \in W_{i_s}, l_s \in W_{i_s}, s = 1, 2) \\
&= \frac{2}{n(p-1)^2} \sum_{i_1=1}^n \sum_{i_2=1}^n \sum_{j \neq l}^n \sigma^2(., x_{2j}) \sigma^2(., x_{2l}) I(j, l \in W_{i_1} \cap W_{i_2}) \\
&= \frac{2}{n(p-1)^2} \sum_{i_1=1}^n \sum_{i_2=1}^n \sum_{j \neq l}^n \sigma^2(., x_{2j}) \left( \sigma^2(., x_{2j}) + O_p\left(\frac{p}{\sqrt{n}}\right) \right) I(j, l \in W_{i_1} \cap W_{i_2}) \\
&= \frac{2}{n(p-1)^2} \sum_{j=1}^n \sigma^4(., x_{2j}) \sum_{i_1=1}^n \sum_{i_2=1}^n \sum_{l \neq j}^n I(j, l \in W_{i_1} \cap W_{i_2}) + O_p\left(\frac{p^2}{n^{1/2}}\right) \\
&= \frac{2}{n(p-1)^2} \sum_{j=1}^n \sigma^4(., x_{2j}) 2(1 + 2^2 + 3^2 + \dots + (p-1)^2) + O_p\left(\frac{p^2}{n^{1/2}}\right) \\
&= \frac{2}{n(p-1)^2} \frac{p(p-1)(2p-1)}{3} \sum_{j=1}^n \sigma^4(., x_{2j}) + O_p\left(\frac{p^2}{n^{1/2}}\right),
\end{aligned} \tag{22}$$

where the third equality follows from Lemma A.2 using the assumption that  $\sigma^2(., x_2)$  is Lipschitz continuous and the second last inequality results from the fact that if  $1 \leq |j_1 - j_2| = s \leq p-1$ , then they are  $(p-s)^2$  pairs of windows whose intersection includes  $j_1$  and  $j_2$ . Taking limits as  $n \rightarrow \infty$  it is seen that

$$E(n^{1/2} \xi'_V A_d \xi_V | X_2 = x_2)^2 \xrightarrow{a.s.} \frac{2(2p-1)}{3(p-1)} E(\sigma^4(., X_2)) = \frac{2(2p-1)}{3(p-1)} \tau^2. \tag{23}$$

From relation (22) it is easily seen that  $E[(\sqrt{n} \xi'_V A_d \xi_V)^2 | \{X_{2j}\}_{j=1}^n]$  remains bounded, and thus  $\text{Var}(n^{1/2} \xi'_V A_d \xi_V)$  also converges to the same limit by the Dominated Convergence Theorem. Hence,  $n^{1/2} \xi'_V A_d \xi_V$  converges in distribution to the designated normal distribution. That the second and third terms in (21) converge in probability to zero are shown in Lemmas C.1, C.2, respectively.  $\square$

### ***Proof of Theorem 2.2. Part 1: Local Additive Alternatives***

Note that we can write  $\hat{\xi}_j = Y_j - \hat{m}_1(\mathbf{X}_{1j})$  as

$$\begin{aligned}
\hat{\xi}_j &= Y_j - m_1(\mathbf{X}_{1j}) - \rho_n \tilde{m}_2(X_{2j}) - [\hat{m}_1(\mathbf{X}_{1j}) - m_1(\mathbf{X}_{1j})] + \rho_n \tilde{m}_2(X_{2j}) \\
&= \xi_j - \Delta_{m_1}(\mathbf{X}_{1j}) + \rho_n \tilde{m}_2(X_{2j}),
\end{aligned} \tag{24}$$

and therefore  $\hat{\xi}_V = \xi_V - \Delta_{m_1V} + \rho_n \tilde{\mathbf{m}}_{2V}$ , where  $\xi_V$ ,  $\Delta_{m_1V}$  and  $\tilde{\mathbf{m}}_{2V}$  are defined as in (10) but using  $\xi_i$ ,  $\Delta_{m_1}(\mathbf{X}_{1i})$  and  $\tilde{m}_2(X_{2i})$ , respectively, instead of  $\hat{\xi}_i$ . Thus, we can write

$$\begin{aligned} \sqrt{n}(MST - MSE) &= \sqrt{n} \hat{\xi}_V' A \hat{\xi}_V = \sqrt{n}(\xi_V - \Delta_{m_1V})' A (\xi_V - \Delta_{m_1V}) + \\ &+ \sqrt{n} 2\rho_n (\xi_V - \Delta_{m_1V})' A \tilde{\mathbf{m}}_{2V} + \sqrt{n} \rho_n^2 \tilde{\mathbf{m}}_{2V}' A \tilde{\mathbf{m}}_{2V}. \end{aligned} \quad (25)$$

By Theorem 2.1,  $\sqrt{n}(\xi_V - \Delta_{m_1V})' A (\xi_V - \Delta_{m_1V}) \xrightarrow{d} N(0, [2p(2p-1)\tau^2]/[3(p-1)])$ . That  $\sqrt{n} 2\rho_n (\xi_V - \Delta_{m_1V})' A \tilde{\mathbf{m}}_{2V} \xrightarrow{p} 0$  and  $\sqrt{n} \rho_n^2 \tilde{\mathbf{m}}_{2V}' A \tilde{\mathbf{m}}_{2V} \xrightarrow{p} a^2 p V(\tilde{m}_2(X_2))$  are shown in Lemma C.3 and Lemma C.4, respectively. This completes the proof of part 1.

## Part 2: Local General Alternatives

Working as in (24) we can write  $\hat{\xi}_V = \xi_V - \Delta_{m_1V} + \rho_{1n} \tilde{\mathbf{m}}_{2V} + \rho_{2n} \tilde{\mathbf{m}}_{12V}$ , where  $\xi_V$ ,  $\Delta_{m_1V}$ ,  $\tilde{\mathbf{m}}_{2V}$  and  $\tilde{\mathbf{m}}_{12V}$  are defined as in (10) but using  $\xi_i$ ,  $\Delta_{m_1}(\mathbf{X}_{1i})$ ,  $\tilde{m}_2(X_{2i})$  and  $\tilde{m}_{12}(\mathbf{X}_{1i}, X_{2i})$ , respectively, instead of  $\hat{\xi}_i$ . Thus  $\sqrt{n}(MST - MSE)$  is

$$\begin{aligned} \sqrt{n} \hat{\xi}_V' A \hat{\xi}_V &= \sqrt{n}(\xi_V - \Delta_{m_1V} - \rho_{1n} \tilde{\mathbf{m}}_{2V})' A (\xi_V - \Delta_{m_1V} - \rho_{1n} \tilde{\mathbf{m}}_{2V}) \\ &+ \sqrt{n} 2\rho_{2n} (\xi_V - \Delta_{m_1V} - \rho_{1n} \tilde{\mathbf{m}}_{2V})' A \tilde{\mathbf{m}}_{12V} + \sqrt{n} \rho_{2n}^2 \tilde{\mathbf{m}}_{12V}' A \tilde{\mathbf{m}}_{12V}. \end{aligned} \quad (26)$$

By Part 1 of the theorem,  $\sqrt{n}(\xi_V - \Delta_{m_1V} - \rho_{1n} \tilde{\mathbf{m}}_{2V})' A (\xi_V - \Delta_{m_1V} - \rho_{1n} \tilde{\mathbf{m}}_{2V})$  converges in distribution to

$$N(a^2 p \text{Var}(m_2(X_2)), [2p(2p-1)\tau^2]/[3(p-1)]).$$

Hence, it is enough to show that  $\sqrt{n} \rho_{2n}^2 \tilde{\mathbf{m}}_{12V}' A \tilde{\mathbf{m}}_{12V} \xrightarrow{p} p b^2 \text{Var}(\tilde{m}_{12}(\mathbf{X}_1, X_2))$  and  $\sqrt{n} 2\rho_{2n} (\xi_V - \Delta_{m_1V} - \rho_{1n} \tilde{\mathbf{m}}_{2V})' A \tilde{\mathbf{m}}_{12V} \xrightarrow{p} 2pab \text{Cov}(\tilde{m}_2(X_2), \tilde{m}_{12}(\mathbf{X}_1, X_2))$ . These are shown in Lemmas C.6 and C.5, respectively.  $\square$

## C Some Detailed Derivations

**Lemma C.1.** *The second term in (21) converges in probability to zero, i.e.*

$$T_{2n} := \sqrt{n} \xi_V' A \Delta_{m_1V} \xrightarrow{p} 0.$$

*Proof.* After some algebra it can be seen that

$$\begin{aligned}
T_{2n} &= \frac{n^{-1/2}(np-1)}{(n-1)p(p-1)} \sum_{i=1}^n \sum_{j \in W_i} \xi_j \sum_{k \in W_i} \Delta_{m_1}(\mathbf{X}_{1k}) \\
&\quad - \frac{n^{-1/2}p}{(n-1)} \sum_{i=1}^n \xi_i \sum_{j=1}^n \Delta_{m_1}(\mathbf{X}_{1j}) - \frac{n^{-1/2}p}{(p-1)} \sum_{i=1}^n \xi_i \Delta_{m_1}(\mathbf{X}_{1i}). \tag{27}
\end{aligned}$$

We will show that each of the three terms above converge in probability to zero conditionally on the set of observed predictors,  $\{\mathbf{X}_j\}_{j=1}^n$ , and thus also unconditionally. Note that, because all windows  $W_i$  are of finite size  $p$ , the first term on the right hand side of (27) can be written as a finite sum of  $p^2$  terms each of which is similar to the last term in (27). Thus, it suffices to show that the last and second terms of (27) converge to zero. For notational simplicity, all expectations and variances in this proof are to be understood as conditional on  $\{\mathbf{X}_j\}_{j=1}^n$ . For the last term in (27) we have

$$\begin{aligned}
n^{-1/2} \sum_{i=1}^n \xi_i \Delta_{m_1}(\mathbf{X}_{1i}) &= n^{-1/2} \sum_{i=1}^n \sum_{j=1}^n w(\mathbf{X}_{1i}, \mathbf{X}_{1j}) (m_1(\mathbf{X}_{1j}) + \xi_j - m_1(\mathbf{X}_{1i})) \xi_i \\
&= n^{-1/2} \sum_{i=1}^n \sum_{j=1}^n w(\mathbf{X}_{1i}, \mathbf{X}_{1j}) (m_1(\mathbf{X}_{1j}) - m_1(\mathbf{X}_{1i})) \xi_i \\
&\quad + n^{-1/2} \sum_{i=1}^n \sum_{j=1}^n w(\mathbf{X}_{1i}, \mathbf{X}_{1j}) \xi_j \xi_i. \tag{28}
\end{aligned}$$

The first term of the right hand side of (28) has zero expectation, so it suffices to show that its variance goes to zero. To this end, we write

$$\begin{aligned}
&\text{Var}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n \sum_{j=1}^n \xi_i w(\mathbf{X}_{1i}, \mathbf{X}_{1j}) (m_1(\mathbf{X}_{1j}) - m_1(\mathbf{X}_{1i}))\right) \\
&= \frac{1}{n} \sum_{i=1}^n \sum_{j_1=1}^n \sum_{j_2=1}^n w(\mathbf{X}_{1i}, \mathbf{X}_{1j_1}) w(\mathbf{X}_{1i}, \mathbf{X}_{1j_2}) \times \\
&\quad \times (m_1(\mathbf{X}_{1j_1}) - m_1(\mathbf{X}_{1i})) (m_1(\mathbf{X}_{1j_2}) - m_1(\mathbf{X}_{1i})) \text{Var}(\xi_i) \\
&\leq \frac{M}{n} \sum_{i=1}^n \sum_{j_1=1}^n \sum_{j_2=1}^n w(\mathbf{X}_{1i}, \mathbf{X}_{1j_1}) w(\mathbf{X}_{1i}, \mathbf{X}_{1j_2}) \times \\
&\quad \times (c \|\mathbf{X}_{1j_1} - \mathbf{X}_{1i}\| c \|\mathbf{X}_{1j_2} - \mathbf{X}_{1i}\|) \\
&= Mc^2 O(\|H_n^{1/2}\|) O(\|H_n^{1/2}\|),
\end{aligned}$$

for some constants  $M$  and  $c$ , where the inequality holds by the assumed conditions for  $m_1(\cdot)$ , and the last equality follows from Lemma A.3. Thus, by the assumptions of Theorem 2.1 the first term of the right hand side of (28) goes in probability to zero. To show that the second term in (28) also goes to 0 in probability since, we will show that its second moment goes to zero. To this end, we write

$$\begin{aligned}
& E \left[ \frac{1}{n} \sum_{i_1=1}^n \sum_{i_2=1}^n \sum_{j_1=1}^n \sum_{j_2=1}^n \xi_{i_1} \xi_{i_2} \xi_{j_1} \xi_{j_2} w(\mathbf{X}_{1i_1}, \mathbf{X}_{1j_1}) w(\mathbf{X}_{1i_2}, \mathbf{X}_{1j_2}) \right] \\
&= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n E(\xi_i^2 \xi_j^2) [w(\mathbf{X}_{1i}, \mathbf{X}_{1j})^2 + w(\mathbf{X}_{1i}, \mathbf{X}_{1i}) w(\mathbf{X}_{1j}, \mathbf{X}_{1j}) \\
&\quad + w(\mathbf{X}_{1i}, \mathbf{X}_{1j}) w(\mathbf{X}_{1j}, \mathbf{X}_{1i})] \\
&\quad + \frac{1}{n} \sum_{i=1}^n E(\xi_i^4) w(\mathbf{X}_{1i}, \mathbf{X}_{1i}) w(\mathbf{X}_{1i}, \mathbf{X}_{1i}) \\
&\leq \frac{M_1^2}{n} \sum_{i=1}^n \sum_{j=1}^n \frac{c^2}{n^2 |H_n| \hat{f}_{\mathbf{X}_1}(\mathbf{X}_{1i})} \left[ \frac{1}{\hat{f}_{\mathbf{X}_1}(\mathbf{X}_{1i})} + \frac{2}{\hat{f}_{\mathbf{X}_1}(\mathbf{X}_{1j})} \right] \\
&\quad + \frac{M_2}{n} \sum_{i=1}^n \frac{c^2}{n^2 |H_n| \hat{f}_{\mathbf{X}_1}(\mathbf{X}_{1i})^2} = O\left(\frac{1}{n|H_n|}\right) + O\left(\frac{1}{n^2|H_n|}\right), \tag{29}
\end{aligned}$$

for some constants  $M_1$ ,  $M_2$  and  $c$ , by the fact that  $\hat{f}_1$  converges uniformly to  $f$  a.s. in the compact support  $S_{\mathbf{X}_1}$  (Ruschendorf 1977). Thus, by the assumptions of Theorem 2.1 the second term of the right hand side of (28) goes in probability to zero.

Consider now the second term in (27). Since  $n^{-1/2} \sum_{i=1}^n \xi_i$  remains bounded in probability, its convergence to zero will follow if we show that  $n^{-1} \sum_{k=1}^n \Delta_{m_1}(\mathbf{X}_{1k}) \xrightarrow{p} 0$ . For later use, we will actually show that

$$\frac{1}{n^{3/4}} \sum_{k=1}^n \Delta_{m_1}(\mathbf{X}_{1k}) = \frac{1}{n^{3/4}} \sum_{k=1}^n (\hat{m}_1(\mathbf{X}_{1k}) - m_1(\mathbf{X}_{1k})) \xrightarrow{p} 0. \tag{30}$$

For this we use (cf. Hansen, 2008)

$$\sup_{\mathbf{x}} |\hat{m}_1(\mathbf{x}) - m_1(\mathbf{x})| = O_p(a_n) \quad \text{where} \quad a_n = \left( \frac{\log n}{n \lambda^{d-1}} \right)^{1/2} + \lambda^2, \tag{31}$$

where  $\lambda \rightarrow 0$  at the same rate as the eigenvalues  $\lambda_i$ ,  $i = 1, \dots, d-1$ , of  $H_n$ . Therefore, the term in the left hand side of (30) is of order

$$\frac{1}{n^{3/4}} n O_p \left( \left( \frac{\log n}{n \lambda^{d-1}} \right)^{1/2} + \lambda^2 \right) = o_p(1),$$

by the assumed conditions stated in (12). This completes the proof of Lemma C.1.  $\square$

**Lemma C.2.** *The third term in (21) converges in probability to zero, i.e.*

$$T_{3n} = \sqrt{n} \Delta'_{m_1 V} A \Delta_{m_1 V} \xrightarrow{p} 0.$$

*Proof.* In this proof we will use  $w_{ij}$  to denote  $w(\mathbf{X}_{1i}, \mathbf{X}_{1j})$ . Writting

$$\begin{aligned} T_{3n} &= \frac{\sqrt{n}(np-1)}{n(n-1)p(p-1)} \sum_{i=1}^n \left( \sum_{j \in W_i} \Delta_{m_1}(\mathbf{X}_{1j}) \right)^2 \\ &\quad - \frac{\sqrt{np}}{n(n-1)} \left( \sum_{i=1}^n \Delta_{m_1}(\mathbf{X}_{1i}) \right)^2 - \frac{\sqrt{np}}{n(p-1)} \sum_{i=1}^n \Delta_{m_1}^2(\mathbf{X}_{1i}), \end{aligned} \quad (32)$$

we have to show that each of the three terms on the right hand side of (32) converges to zero in probability. First notice that, because all windows  $W_i$  are of finite size  $p$ , the first term on the right hand side of (32) can be written as a finite sum of  $p^2$  terms each of which is similar to the last term in (32). Therefore, to show that the first and third terms in (32) go to zero in probability it is enough to show that  $n^{-1/2} \sum_{i=1}^n \Delta_{m_1}^2(\mathbf{X}_{1i}) \xrightarrow{p} 0$ . Using (31), it is easy to see that

$$n^{-1/2} \sum_{i=1}^n \Delta_{m_1}^2(\mathbf{X}_{1i}) \leq n^{1/2} O_p(a_n)^2 = o_p(1).$$

That the second term on the right hand side of (32) converges in probability to zero follows directly from (30).  $\square$

**Lemma C.3.** *The second term in (25) converges in probability to zero, i.e.*

$$\sqrt{n} 2\rho_n (\boldsymbol{\xi}_V - \Delta_{m_1 V})' A \tilde{\mathbf{m}}_{2V} \xrightarrow{p} 0.$$

*Proof.* By the definition of the matrix  $A$ , we can write  $(\boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1 V})' A \tilde{\mathbf{m}}_{2V}$  as

$$\begin{aligned} & \frac{np-1}{n(n-1)p(p-1)} \sum_{i=1}^n \left[ \sum_{j=1}^n \tilde{m}_2(X_{2j}) I(j \in W_i) \right] \left[ \sum_{k=1}^n (\xi_k - \Delta_{m_1}(\mathbf{X}_{1k})) I(k \in W_i) \right] \\ & - \frac{1}{n(n-1)p} \left[ p \sum_{i=1}^n \tilde{m}_2(X_{2i}) \right] \left[ p \sum_{i=1}^n (\xi_i - \Delta_{m_1}(\mathbf{X}_{1i})) \right] \\ & - \frac{p}{n(p-1)} \sum_{i=1}^n \tilde{m}_2(X_{2i}) (\xi_i - \Delta_{m_1}(\mathbf{X}_{1i})). \end{aligned}$$

Using Lemma A.2 and the fact that  $\tilde{m}_2(\cdot)$  is Lipschitz continuous, the sum in the first term can be expressed as

$$\begin{aligned} & p \sum_{i=1}^n [\tilde{m}_2(X_{2i}) + O(n^{-1/2})] \left[ \sum_{k=1}^n (\xi_k - \Delta_{m_1}(\mathbf{X}_{1k})) I(k \in W_i) \right] \leq \\ & \leq p \sum_{k=1}^n \left[ \sum_{i=1}^n \tilde{m}_2(X_{2i}) I(i \in W_k) \right] (\xi_k - \Delta_{m_1}(\mathbf{X}_{1k})) + p^2 O(n^{-1/2}) \sum_{k=1}^n |(\xi_k - \Delta_{m_1}(\mathbf{X}_{1k}))| \\ & = p^2 \sum_{k=1}^n \tilde{m}_2(X_{2k}) (\xi_k - \Delta_{m_1}(\mathbf{X}_{1k})) + O_p(p^2 n^{1/2}), \end{aligned}$$

so that

$$\begin{aligned} \sqrt{n} \rho_n \tilde{\mathbf{m}}'_{2V} A (\boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1 V}) &= \frac{an^{1.25}p}{n-1} \left[ \frac{1}{n} \sum_{i=1}^n \tilde{m}_2(X_{2i}) (\xi_i - \Delta_{m_1}(\mathbf{X}_{1i})) \right] \\ &\quad - \frac{an^{1.25}p}{n-1} \left[ \frac{1}{n} \sum_{i=1}^n \tilde{m}_2(X_{2i}) \right] \left[ \frac{1}{n} \sum_{i=1}^n (\xi_i - \Delta_{m_1}(\mathbf{X}_{1i})) \right] + O_p\left(\frac{1}{n^{1/4}}\right). \end{aligned}$$

Using the fact that  $E(\tilde{m}_2(X_{2i})) = E(\tilde{m}_2(X_{2i})\xi_i) = E(\xi_i) = 0$ , relation (30) and also that  $n^{-3/4} \sum_{i=1}^n \tilde{m}_2(X_{2i}) \Delta_{m_1}(\mathbf{X}_{1i}) \xrightarrow{p} 0$ , as is shown in a similar way to (30), completes the proof of the lemma.  $\square$

**Lemma C.4.** *The third term in (25) converges in probability to  $a^2 p V(\tilde{m}_2(X_2))$ , i.e.*

$$\sqrt{n} \rho_n^2 \tilde{\mathbf{m}}'_{2V} A \tilde{\mathbf{m}}_{2V} \xrightarrow{p} a^2 p V(\tilde{m}_2(X_2)).$$

*Proof.* Writing

$$\begin{aligned} \tilde{m}'_{2V} A \tilde{m}_{2V} &= \frac{np}{n-1} \left\{ \left[ \frac{1}{n} \sum_{i=1}^n \tilde{m}_2^2(X_{2i}) \right] - \left[ \frac{1}{n} \sum_{i=1}^n \tilde{m}_2(X_{2i}) \right]^2 \right\} + O\left(\frac{1}{n^{1/2}}\right) \\ &= p \{ E \tilde{m}_2^2(X_2) - [E \tilde{m}_2(X_2)]^2 \} + O_p\left(\frac{1}{n^{1/2}}\right), \end{aligned}$$

it follows that

$$\sqrt{n}\rho_n^2 \tilde{m}'_{2V} A \tilde{m}_{2V} = a^2 p \text{Var}(\tilde{m}_2(X_2)) + O_p\left(\frac{1}{n^{1/2}}\right),$$

which completes the proof.  $\square$

**Lemma C.5.** *The second term in (26) converges in probability to  $2pab\text{Cov}(\tilde{m}_2(X_2), \tilde{m}_{12}(\mathbf{X}_1, X_2))$ , i.e.*

$$\sqrt{n}2\rho_{2n}(\boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1V} - \rho_{1n}\tilde{\mathbf{m}}_{2V})' A \tilde{\mathbf{m}}_{12V} \xrightarrow{p} 2pab\text{Cov}(\tilde{m}_2(X_2), \tilde{m}_{12}(\mathbf{X}_1, X_2)).$$

*Proof.* By the definition of the matrix  $A$ , we can write

$$\begin{aligned} \sqrt{n}\rho_{2n}(\boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1V} - \rho_{1n}\tilde{\mathbf{m}}_{2V})' A \tilde{\mathbf{m}}_{12V} &= \sqrt{n}\rho_{2n} \frac{np-1}{n(n-1)p(p-1)} \times \\ &\times \sum_{i=1}^n \left[ \sum_{j=1}^n \tilde{m}_{12}(\mathbf{X}_{1j}, X_{2j}) I(j \in W_i) \right] \left[ \sum_{k=1}^n (\xi_k - \Delta_{m_1}(\mathbf{X}_{1k}) - \rho_{1n}\tilde{m}_2(X_{2k})) I(k \in W_i) \right] \\ &- \sqrt{n}\rho_{2n} \frac{1}{n(n-1)p} \left[ p \sum_{i=1}^n \tilde{m}_{12}(\mathbf{X}_{1i}, X_{2i}) \right] \left[ p \sum_{i=1}^n (\xi_i - \Delta_{m_1}(\mathbf{X}_{1i}) - \rho_{1n}\tilde{m}_2(X_{2i})) \right] \\ &- \sqrt{n}\rho_{2n} \frac{p}{n(p-1)} \sum_{i=1}^n \tilde{m}_{12}(\mathbf{X}_{1i}, X_{2i}) (\xi_i - \Delta_{m_1}(\mathbf{X}_{1i}) - \rho_{1n}\tilde{m}_2(X_{2i})). \end{aligned} \quad (33)$$

Noting that  $n^{-3/4} \sum_{i=1}^n \xi_i \tilde{m}_{12}(\mathbf{X}_{1i}, X_{2i}) \xrightarrow{p} 0$ , and  $\frac{1}{n^{3/4}} \sum_{i=1}^n \Delta_{m_1}(\mathbf{X}_{1i}) \tilde{m}_{12}(\mathbf{X}_{1i}, X_{2i}) \xrightarrow{p} 0$ , which follows by arguments similar to (30), the third term in (33) goes in probability to  $[pab/(p-1)]E(\tilde{m}_2(X_2)\tilde{m}_{12}(\mathbf{X}_1, X_2))$ . Also, using (30), and the facts  $E(\tilde{m}_{12}(\mathbf{X}_1, X_2)) = 0$ , and  $n^{-3/4} \sum_{i=1}^n \xi_i = o_p(1)$ , the second term in (33) goes to  $pabE(\tilde{m}_2(X_2))E(\tilde{m}_{12}(\mathbf{X}_1, X_2))$  in probability. Next, the component of the first term in (33) that corresponds to

$$\sum_{j=1}^n \sum_{k=1}^n \sum_{i=1}^n \tilde{m}_{12}(\mathbf{X}_{1j}, X_{2j}) (\xi_k - \Delta_{m_1}(\mathbf{X}_{1k})) I(j \in W_i) I(k \in W_i)$$

goes to zero in probability by arguments similar to those used for the last term in (33). Set

$\bar{m}_2^i(X_{2i}) = \frac{1}{p} \sum_{j=1}^n \tilde{m}_2(X_{2j}) I(j \in W_i)$  and  $\bar{m}_{12}^i(\cdot, X_{2i}) = \frac{1}{p} \sum_{j=1}^n \tilde{m}_{12}(\mathbf{X}_{1j}, X_{2j}) I(j \in W_i)$ , so

that

$$\begin{aligned}\frac{1}{p} \sum_{j=1}^n \tilde{m}_{12}(\mathbf{X}_{1j}, X_{2j}) I(j \in W_i) &= \bar{m}_{12}^i(\cdot, X_{2i}) + o_p(1), \\ \frac{1}{p} \sum_{j=1}^n \tilde{m}_2(X_{2j}) I(j \in W_i) &= \bar{m}_2^i(X_{2i}) + o_p(1).\end{aligned}$$

The remaining component of the first term in (33) can be written as

$$\begin{aligned}& \frac{(np-1)ab}{n(n-1)p(p-1)} \sum_{j=1}^n \sum_{k=1}^n \sum_{i=1}^n \tilde{m}_{12}(\mathbf{X}_{1j}, X_{2j}) \tilde{m}_2(X_{2k}) I(j \in W_i) I(k \in W_i) \\ &= \frac{(np-1)pab}{(n-1)(p-1)n} \frac{1}{n} \sum_{i=1}^n \bar{m}_{12}^i(\cdot, X_{2i}) \bar{m}_2^i(X_{2i}) + o_p(1) \xrightarrow{p} \frac{p^2 b^2}{p-1} E[\tilde{m}_{12}(\mathbf{X}_1, X_2) \tilde{m}_2(X_2)],\end{aligned}$$

completing the proof.  $\square$

**Lemma C.6.** *The third term in (26) converges in probability to  $pb^2 \text{Var}(\tilde{m}_{12}(\mathbf{X}_1, X_2))$ , i.e.*

$$\sqrt{n} \rho_{2n}^2 \tilde{\mathbf{m}}'_{12V} A \tilde{\mathbf{m}}_{12V} \xrightarrow{p} pb^2 \text{Var}(\tilde{m}_{12}(\mathbf{X}_1, X_2)).$$

*Proof.* Note that we can write  $\sqrt{n} \rho_{2n}^2 \tilde{\mathbf{m}}'_{12V} A \tilde{\mathbf{m}}_{12V}$  as

$$\begin{aligned}& \frac{(np-1)b^2}{n(n-1)p(p-1)} \sum_{i=1}^n \left[ \sum_{j=1}^n \tilde{m}_{12}(\mathbf{X}_{1j}, X_{2j}) I(j \in W_i) \right] \left[ \sum_{k=1}^n \tilde{m}_{12}(\mathbf{X}_{1k}, X_{2k}) I(k \in W_i) \right] \\ & - \frac{pb^2}{n(n-1)} \left[ \sum_{i=1}^n \tilde{m}_{12}(\mathbf{X}_{1i}, X_{2i}) \right] \left[ \sum_{i=1}^n \tilde{m}_{12}(\mathbf{X}_{1i}, X_{2i}) \right] \\ & - \frac{pb^2}{n(p-1)} \sum_{i=1}^n \tilde{m}_{12}(\mathbf{X}_{1i}, X_{2i})^2.\end{aligned}\tag{34}$$

Clearly, the third term in (34) goes to  $[pb^2/(p-1)]E[\tilde{m}_{12}(\mathbf{X}_1, X_2)^2]$  in probability, and the second term in (34) goes to  $pb^2[E(\tilde{m}_{12}(\mathbf{X}_1, X_2))]^2$  in probability. Using the same notation as in lemma C.5, the first term in (34) is equal to

$$\frac{(np-1)pb^2}{n(n-1)(p-1)} \sum_{i=1}^n [\bar{m}_{12}^i(\cdot, X_{2i})]^2 + o_p(1) \xrightarrow{p} \frac{p^2 b^2}{p-1} E[(\tilde{m}_{12}(\mathbf{X}_{1i}, X_{2i}))^2],$$

completing the proof.  $\square$

## References

- [1] At-Sahalia, Y., Bickel, P. J. and Stoker, T.M. (2001). Goodness-of-fit tests for kernel regression with an application to option implied volatilities. *Journal of Econometrics*, 105, 363-412.
- [2] Abramovich, F., Benjamini, Y., Donoho, D.L. and Johnstone, I.M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, 34, 584-653.
- [3] Akritas, M. G. and Papadatos, N. (2004). Heterocedastic One-Way ANOVA and Lack-of-Fit Tests. *Journal of the American Statistical Association*, 99, Theory and Methods.
- [4] Benjamini, Y.; Gavrilov, Y. (2009). A Simple Forward Selection Procedure Based on False Discovery Rate Control. *The Annals of Applied Statistics*, 3, 179-198.
- [5] Benjamini, Y.; Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57 (1): 289-300.
- [6] Benjamini, Y., Krieger, A.M., Yekutieli, D. (2006). Adaptive Linear Step-up False Discovery Rate controlling procedures. *Biometrika*, 93 (3): 491-507.
- [7] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29, 1165-1188.
- [8] Birge, L. and Massart, P. (2001) A generalized Cp criterion for Gaussian model. *Technical report*, Lab. De Probabilities, Univ. Paris VI. (<http://www.proba.jussieu.fr/mathdoc/preprints/index.html#2001>)
- [9] Bunea, F., Wegkamp, M. and Auguste, A. (2006). Consistent variable selection in high dimensional regression via multiple testing. *Journal of Statistical Planning and Inference*, 136, 4349-4364.
- [10] Candès, E., and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35, 2313-2351.
- [11] Delgado, M. A. and Manteiga, W. G. (2001). Significance Testing in Nonparametric Regression Based on the Bootstrap *The Annals of Statistics*, 29, 1469-1507.
- [12] Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81, 425-555.
- [13] Dvoretzky, A.; Kiefer, J. and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27 (3), 642-669.
- [14] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression. *The Annals of Statistics*, 32, 407-499.

- [15] Fan, J. and Jiang, J.(2005). Nonparametric Inferences for Additive Models. *Journal of the American Statistical Association*, 100, 890-907.
- [16] Fan, Y. and Li, Q. (1996). Consistent model specification tests: Omitted variables and semiparametric functional forms. *Econometrica*, 64 (4), 865-890.
- [17] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348-1360.
- [18] Fan, J., Zhang, C. M., and Zhang, J. (2001), Generalized Likelihood Ratio Statistics and Wilks Phenomenon. *The Annals of Statistics*, 29, 1531-1593.
- [19] Foster, D. P. and Stine, R. A. (2004). Variable selection in data mining: building a predictive model for bankruptcy. *Journal of the American Statistical Association*, 99, 303-313.
- [20] Hansen, B. E., (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24, 726-748.
- [21] Hart, J. D. (1997). *Nonparametric Smoothing and Lack of fit tests*, Springer.
- [22] Horowitz, J.L. and Mammen, E. (2004). Nonparametric estimation of an additive model with a link function. *The Annals of Statistics*, 32, 2412-2443.
- [23] Huang, J., Horowitz, J. L., and Wei, F. (2010). Variable Selection in Nonparametric Additive Models. Available at <http://faculty.wcas.northwestern.edu/~jlh951/papers/HHW-npam.pdf>
- [24] Lavergne, P. and Vuong, Q. (2000). Nonparametric Significance Testing. *Econometric Theory*, 16, 576-601.
- [25] Li, K. C., (1991). Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association*, 86, 316-327.
- [26] Li, R. and Liang, H. (2008). Variable selection in Semiparametric Regression Modeling. *The Annals of Statistics*, 36, 261-286.
- [27] Linton, O. and Nielsen, J.P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, 82, 93-100.
- [28] Mammen, E., Linton, O. and Nielsen, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *The Annals of Statistics*, 27, 1443-1490.
- [29] Newey, W. K. (1994). Kernel estimation of partial means. *Econom. Theory*, 10, 233-253.
- [30] Racine, J., Hart, J.D. and Li, Q. (2006). Testing the significance of categorical predictor variables in nonparametric regression models. *Econometric Reviews*, 25, 523-544.
- [31] Rice, J. (1984). Bandwidth choice for nonparametric regression. *The Annals of Statistics*, 12, 1215-1230.

- [32] Ruschendorf, L. (1977). Consistency of Estimators for Multivariate Density Functions and For the Mode. *Sankhya: The Indian Journal of Statistics*, 39, 243-250.
- [33] Stone, C. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10, 1040-1053.
- [34] Stone, C. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, 13, 689-705.
- [35] Storlie, C. B, Bondell, H. D, Reich, B. J, Zhang, H. H. (2011). Surface Estimation, Variable Selection, and the Nonparametric Oracle Property. *Statistica Sinica*, 21(2), 679-705.
- [36] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58, 267-288.
- [37] Tibshirani, R. and Knight, K. (1999). The covariance inflation criterion for adaptive model selection. *Journal of the Royal Statistical Society B*, 61, 529-546.
- [38] Wang, H. and Xia, Y. (2008). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association*, 104, 747-757.
- [39] Wang, L., Akritas, M. G. and Keilegom, I.V. (2008). An ANOVA-type Nonparametric Diagnostic Test for Heterocedastic Regression Models. *Journal of Nonparametric Statistics*, 20, 365-382.
- [40] Zou, H. (2006). The Adaptive Lasso and its Oracle Properties. *Journal of the American Statistical Association*, 101(476), 1418-1429.